网络资源保存的智慧化转型研究

张学青

（国家图书馆，北京　100081）

［摘要］研究网络资源保存现存的问题及面临的挑战,深入分析网络资源保存项目现阶段发展状况与发展难点，按照智慧图书馆建设的相关要求，从资源建设、数据加工管理、对外服务等层面 提出整个项目智慧化转型的思路与具体方法。

［关键词］网络资源保存；　智慧化转型；　智慧化服务；　知识服务

［中图分类号］G250.73

0 引言

网络资源保存相关项目在国内的发展已有近二十年,这期间互联网发生了很多根本性的变化，网络对人们生产生活的影响越来越深入，人们和网络之间的联系越来越来紧密，原有的网络资源保存的一些理念与做法与新的网络发展形势已经不相适应。同时，数字图书馆建设也步入“智慧化”时代，智慧图书馆建设从理论探索逐步向实践迈进，对图书馆的各项工作提出很多新的要求，也提供了不少新的机遇。以此为契机，应用智慧图书馆的资源建设与对外服务等理念，重新审视网络资源保存发展过程中出现的一些问题，促进网络资源保存的智慧化转型，以期更好的发挥网络资源的作用，深入挖掘其隐藏价值，让网络资源保存类项目更深入地融入智慧图书馆建设的大潮中，是网络资源保存相关项目发展的必由之路，也是目前的最佳选择。

1 网络资源保存项目发展现状

网络资源是数字资源重要的组成部分，进入信息时代以后，人们逐渐认识到网络信息中蕴含的巨大价值。早在上世纪九十年代末，英美等一些发达国家就开始着手启动了一批网络信息保存的相关项目。从 2003 年开始，国家图书馆也启动了自己的网络信息资源保存项目，开始了网络资源收集、保存等工作。国家图书馆在 2009 年成立了互联网信息资源保存保护中心，开始了这项工作的深入研究。2014 年借助数字图书馆推广工程，全国很多地方馆也加入到了这项工作中。

经过近 20 年的持续研究与探索，在国家图书馆的引领下，全国图书馆界建立了较为完整的自上而下的网络资源保存体系，保存范围与保存能力都有所提升。保存范围包括政府公开信息、国内外重要网站、重大专题等内容。采集方式上，国家图书馆利用虚拟化技术提升采集效率，建设云共享式网络资源采集和保存平台，支持国家图书馆与多个地方馆开展共享式、分布式、协同式业务，解决了技术力量不足的地方馆开展工作遇到的问题。针对近年来网络资源产生的新变化，国家图书馆持续进行技术更新，创新采集与管理的算法与策略，实现了增量采集增量回放等功能，并实现了视频资源的高效采集与流畅回放，进一步适应了互联网资源的移动化视频化趋势。到 2021 年初，国家图书馆采集与保存的网络资源数据量已超过 300TB。[1]

各级地方馆也根据自身的实际情况，开展了具有当地特色的网络资源保存相关工作。比如 2009 年开始的中国政府公开信息整合项目，在国家图书馆的带领与推广下，到 2019 年底已经有超 300 个全国各级图书馆参与其中，实现了对各级政府公开信息的收集、整理、保存、开发与利用，在促进全国公共图书馆政府信息服务式方面取得了良好的成效。

2 网络资源保存现存的问题与挑战

经过近二十年的发展，网络资源保存相关工作在保存的数据量、保存范围、参与的主体数量等方面都有了长足的进步。但随着项目不断深入，规模不断扩大及网络的客观条件变化，也逐渐暴露了一些问题。

2.1 网络发展的新阶段带来全新的要求

近二十年中国内网络发生了很多根本性的变化，有些变化之彻底是所有人都始料未及的。2008 年 6 月底，中国网民数量达 2.53 亿，网民规模跃居世界第一，其中手机网民规模达到 7305 万人，中国互联网开始快速前进的步伐。然而，网民普及率只有 19.1%，低于当时全球平均水平 21.1%，网民主体仍旧是 30 岁及以下的年轻群体，网络主要集中在 PC 端，移动网络刚开始兴起，电子商务也处于初创阶段，远未成熟。[2]而到了 2023 年 6 月，我国网民规模到达 10.79 亿，互联网普及率达 76.4%，城乡上网差距缩小，农村网民规模达 3.01 亿人，移动通信方面更是发展迅速，使用手机上网的比例达 99.8%。即时通信、网络视频、短视频等个人社交自媒体继续快速发展，用户使用率分别 97.1%、96.8%、95.2%。[3]现在的网络和 20 年前的网络几乎是完全不同的两个概念，无论从规模还是从质量上，还是对人们日常生活的影响上都发生了翻天覆地的变化，网络资源的丰富、复杂程度早已今非昔比，加上网络资源保存项目本身就是在探索中前进，很多条件并不成熟，国内网络后来的发展又超出很多人的意料，旧的标准逐渐难以适应新的形势，以当时的理念发展设计的网络资源采集项目，不能适应当前的互联网发展状态也确实算不上意外。

2.2 系统管理效率不高 采集方式不够智能

随着 5G 网络的深入发展网络资源的海量、多来源、异构等特点越来越明显。移动网络、自媒体相关内容所占比重及影响也越来越大，这些变化都是前所未有的。现行的系统对日益增长的海量数据已经越来越难以高效管理，存储与维护的压力都非常大，更缺少对移动网络、自媒体音视频等非传统网络资源的处理手段，无法完全适应现阶段网络资源发展的要求。传统的爬虫采集资源方式智能化程度较低，经常需要较多的人工干预，尤其是选择性采集专题信息更是如此，也不能有效处理新兴网络资源。

2.3 服务方式单一，资源利用率有待提升

在采集数据不断增长的同时，网络资源保存项目的服务方式多数还停留在比较传统的关键词检索与直接展示阶段，即使是这种揭示方式在数据逐渐增大时阅览体验也不够好。作为一种知识密度较低的数字资源，如果不能充分发掘其中的隐藏价值，大部分网络资源只能停留在存储的状态，无法发挥馆藏的作用。

2.4 人才培养不足

网络保存项目的持续发展需要一批不仅熟悉相关业务流程，而且在网络技术、数字资源建设与服务等方面有过硬的素质，并有保持对新兴技术的学习热情与能力的工作人员。智慧图书馆时代开展智慧服务以馆员执行为核心，目前人才队伍建设方面很多工作做的不够。

3 智慧化的背景与具体措施

3.1 智慧化的必要性与可行性

网络资源保存项目可持续发展，关键在提高资源资源利用效率与共享程度、提升服务质量全方位多角度满足用户需求，这正是智慧化的优势。

2021 年 3 月《中华人民共和国国民经济发展第十四个五年规划和 2035 年远景目标纲要》提出了发展智慧图书馆的战略，将图书馆的智慧化转型与数字中国的国家战略结合在了一起。[4]同年 9 月，国家图书馆正式发布《国家图书馆“十四五”发展规划》，规划中提出了图书馆智慧转型的任务和建设全国智慧图书馆体系的构想，以满足智慧化的新诉求，推动我国图书馆事业由数字化向智慧化发展。[5]

当下，智慧图书馆建设由理论研究逐渐进入实践实施阶段，借助这一形势完成智慧化转型，并在转型过程中解决发展中的问题，把网络保存项目自身的建设融入智慧图书馆建设的轨道中，提升智慧图书馆建设的质量，反过来也提升网络资源保存自身的发展水平，是网络保存项目现阶段发展的必然选择。

随着硬件技术的高速发展 PB 级数据存储与计算已经不是问题，受此影响，软件方面在算法、大数据处理、计算力等领域也有了长足的发展。人工智能相关技术进入了认知智能时代，开始真正解决问题，并带了不少实际价值。比如在自然语言处理上，人工智能系统在语法分析、答案寻找、语音识别等方面已经越来越接近或超过人类的水平。[6]目前资源联建的形势可以使有条件率先开展的图书馆的一些成功的经验较容易推广到全国地方馆施行。所以目前无论从形势政策上还是技术条件上，实现项目的智慧化转型都是一个比较合适的时机。

3.2 智慧化的具体措施

网络资源保存的智慧化最终要实现项目全流程的智慧化，其中采集智慧化是基础与必要准备；资源建设智慧化是主要手段；对外服务智慧化是最终目的和智慧化结果的直接体现；各项新兴技术的充分合理应用是智慧化转型的基础；培训一支素质过硬的馆员人才队伍是整个工作得以可持续进行的保障。需要说明的是，各个流程的智慧化是有轻重缓急的，并不是也不能完全按照工作顺序进行。在条件有限时应优先以智慧化服务这一个目标为导向，调整资源建设的规划。

3.2.1 资源采集与管理智慧化

在当前的采集工具基础上，增加采集的智能性，尤其是目标选择和资源审核检验等环节减少人工干预的程度，提高工作效率。

智慧图书馆的信息组织对象是以知识资源为核心的数据馆藏，需要对各类馆藏资源做数据化的知识元加工处理，通过对知识元的揭示，重新构建出一个多维度互联的知识系统。用户可以直接从这个系统中获取知识，从而获得真正的知识服务、智慧服务。[7] 同理，资源建设是网络资源保存智慧转型的重点，细粒度、知识化的网络资源是整个项目智慧化的基石。

资源建设的转型应该从项目的实际情况出发，比如技术条件不具备时可以先从部分类型数据和部分资源内容开始，然后纵深发展。比如先处理组织良好、来源权威的结构化数据，再处理半结构化数据，技术与经验成熟后再着手一般化的非结构化资源的建设。资源内容上也可以先选取内容质量规律良好的部分比如网络专题、政府公报等。国家图书馆已经基于保存的政府信息，政府公报等网络资源，抽取其中相关知识，作了智慧化的尝试。图 1 所示即为基于政府公报知识建设成的机构图谱，比较直观的展示了各机构间的关系。智慧化的资源选择可以从局部逐渐延伸到整体，从参与的主体上，也可以由少数条件具备的馆根据自身情况先行建设，然后带动更多的建设单位参与进来，最终有利于节省成本推广经验。

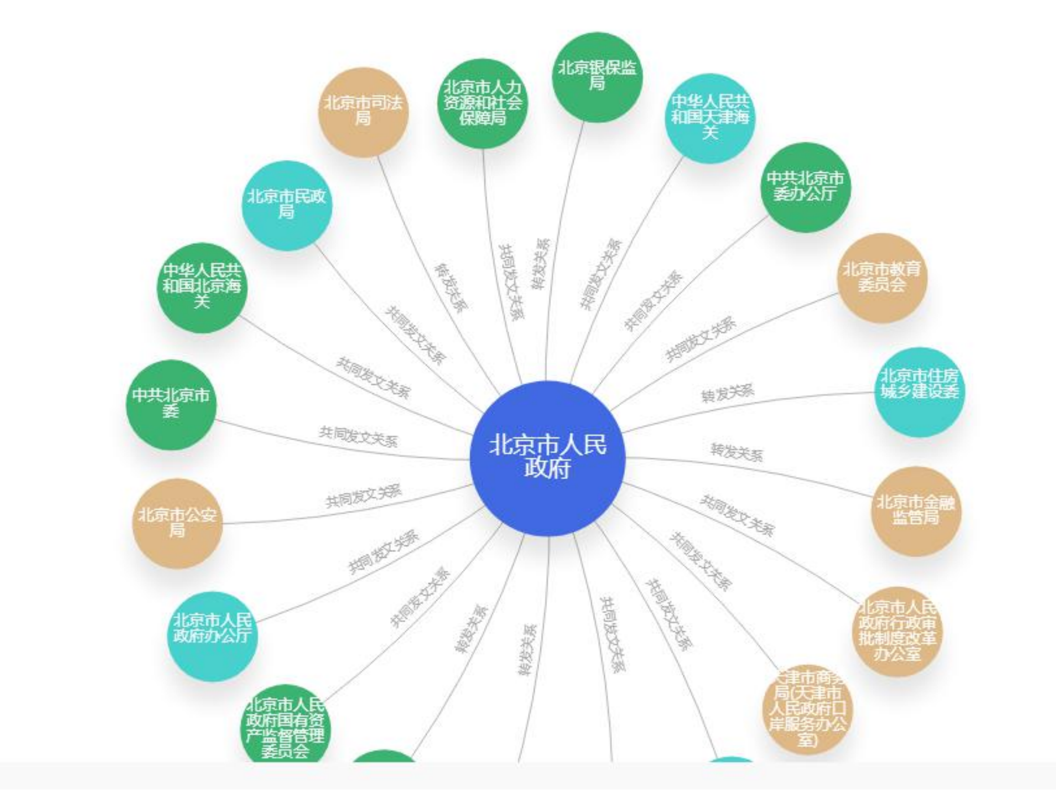


图 1 基于政府公开信息创建的机构图谱

3.2.2 对外服务智慧化

在大量细粒度标识，知识化存储的网络资源基础上，开展深层次的知识服务、智慧服务，是智慧化转型的最直观体现也是最终目标，也是需要持续进行的工作。与以往的被动、单一服务方式不同，转型后可以提供诸如智能问答、知识查询、知识可视化、个性化推送等新型服务形式。并且可以和其它系统交流共享知识资源，在知识共享层面上打通和外界的关联路径，联通不同地域的数据孤岛，构建成更大范围的知识网络、智慧网络。在用户画像等外部知识的配合下，增加服务的主动性，具体性，发展以用户需求为导向，满足用户需求引导用户需求创造用户需求的全新服务方式。服务方式的智慧化，将彻底打通由数据到信息，由信息到知识，由知识到智能，由智能到智慧的数字资源建设通道，这条通道是智慧图书馆时代资源建设的主流模式，不仅能使网络资源的价值大幅提升也可为其它数字资源甚至传统资源的资源建设提供思路。

3.2.3 智慧化的人才队伍培养

智慧图书馆建设的核心并不是先进技术的运用和智慧环境的搭建，而是馆员能力建设。[8]智慧图书馆时代对馆员的能力又有了很多新的要求，智慧化的程度与质量是和参与其中的馆员素质紧密结合在一起的，智慧化的人才队伍甚至本身就是智慧化的一部分。具体到网络资源保存本身，应该在用户需求分析能力、数据管理服务能力、 新技术研发运用能力、智慧服务能力等方面强化和完善，并最终构建专业核心竞争体系。

智慧图书馆建设任务的提出是基于用户需求的，精准把握用户需求是智慧服务需要不断满足的要求。网络资源保存的馆员要从网络资源本身出发去了解、挖掘用户的需求，并能根据网络资源的特点和不同用户的特征提供有针对性的个性化服务。

智慧化转型的最直接体现在于对外提供智慧服务，而数据是一切形式服务的基础。网络资源保存的馆员要加强数据素养，加大网络资源采集尤其是开发的力度，将网络资源包括的文本类、图像音视频类、结构化、非结构化各类资源进行知识组织与揭示，挖掘其中隐藏价值，并最终转化为服务的能力，提高服务质量。

智慧服务需要的技术是动态变化的，具体到网络资源就更为复杂，由于直接面对网络前沿，与数据、信息、网络、通讯、存储计算等相关的各项技术都会对网络资源生命周期的各个流程产生影响，任何技术上的变革与更新都会直接影响网络资源本身，这就要求馆员要有持续的学习能力，不断提升自己，学习各种技术的研发与运用，充分了解各种技术的性能与特点，并能选择其中合适的技术来实现智慧服务的目标。

智慧化转型后需要能够对外提供精准快捷深入专业的、以数据分析与知识发现为主要特征的智能与智慧服务。这要求从事网络资源保存的馆员能够在现有技术系统支持下，根据用户的需求，运用自己甚至团队的专业知识，满足用户个性化的复杂的深度知识需求，智慧服务是智慧化的重要特征，更是智慧馆员价值与能力的集中体现。

项目智慧化转型并最终提供智慧化服务是需要不断探索的工作，需要持续的人才投入，在强化和完善各种能力的基础上，构建馆员专业核心竞争力体系，深入开展智慧馆员培训工作，培养一批具有专业核心竞争力体系的网络资源保存人才队伍，随时充分挖掘网络资源隐藏价值，深度利用所保存的网络资源，实现信息增值，是网络资源保存智慧化转型并可持续发展的关键。

3.3 建立成熟的智慧化资源建设技术体系

网络资源保存智慧化的远景目标，应该是建立一套成熟的资源建设技术体系，这套体系形成后，可以让后续采集到的各种网络资源，经过智慧化建设后，直接提供深层次、多方位的智慧服务，如图 2 所示。

这套技术体系以深度加工的网络资源为基础，以当前的人工智能的各项技术为驱动，最终将把网络资源的服务水平提升到一个新的高度。这套体系的流畅运行不是孤立的，必须有一支专业的人才队伍做为保障。

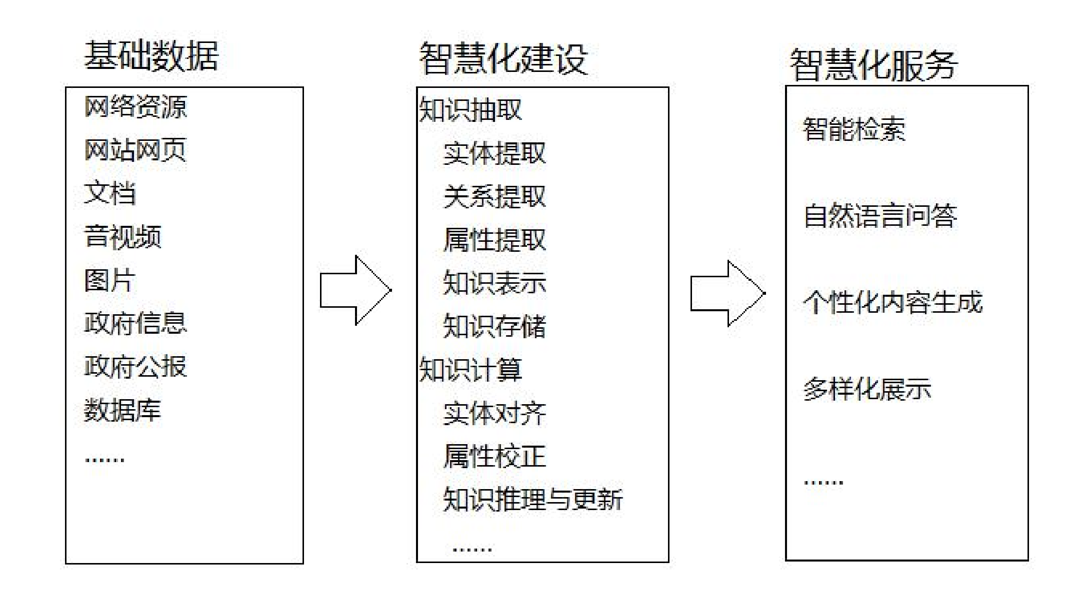


图 2 网络资源智慧化技术体系

智慧图书馆建设仍在探索中发展，很多概念、标准不够明晰，智慧化用到的各项技术发展程度不一，有些技术没有达到实用的程度或者代价昂贵，难以实际应用。网络资源本身数量巨大、种类繁多且信息密度较低，智慧化转型涉及的技术复杂多变，新技术的成熟运用是一个过程，缺少现成的方案可供参考，需要工作人员在实际工作中不断地积累总结经验。

进入 5G 时代后，网络资源的组成更为复杂，无论是从数量、类型上还是从传播的速度更新的频率以及对社会生活的影响上都和之前有根本不同，每一项变化都给网络资源保存带来了直接的挑战。参与网络资源的主体要根据网络实际情况与特点来调整工作方向与对策。

从实际情况看，目前参与网络资源保存建设的主体是国家图书馆带领下的各级地方图书馆，客观条件不一，人员配置不一。普遍存在硬件条件不足，具有相关专业知识的人才不足的问题，需要在国家图书馆的带领下，作好人员培训与经验推广，并且考虑参与项目馆的自身实际，选择开展力所能及的转型方向和范围来开展工作。

4 小结

网络资源保存智慧化转型的最终目标，是提高网络资源的利用效率，让网络资源成为馆藏的有机组成部分，把网络资源保存相关项目的发展融入智慧化建设的大潮中，智慧化转型是长期的逐渐的，需要做好应对各种挑战的准备。本文总结了这一过程的实现步骤和具体措施，希望能为网络资源保存智慧化转型提供指导，也可为其它类型数字图书馆项目的智慧化转型提供思路上的借鉴。

[参考文献]

[1]魏大威，季士妍. 国家图书馆网络信息资源采集与保存平台关键技术实现[J]. 图书馆，2021(3):45-50.

[2] 中国互联网络信息中心. 第 22 次《中国互联网络调查统计报告》[EB/OL].[2023-11-26]http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/201206/t20120612_26713.htm.

[3]中国互联网络信息中心. 第52次中国互联网络调查统计报告［DB/OL］.［2023-11-26］https://www.cnnic.net.cn/n4/2023/0828/c88-10829.html
[4]中华人民共和国国民经济和社会发展第十四个五年规划和2035年远景目标纲要[EB/OL].[2023-11-26] http://www.gov.cn/xinwen/2021-03/13/content_5592681.htm
[5]饶权.全国智慧图书馆体系：开启图书馆智慧化转型新篇章[J].中国图书馆学报，2021(1):4-14.
[6]杨正洪，郭良越， 刘玮. 人工智能与大数据技术导论[M]. 北京：清华大学出版社, 2019：5.
[7]徐婷.从数字化到数据化：信息组织工作如何应对图书馆智慧化转型[J].图书馆界，2022(2):57-61.
[8]初景利，张国瑞.面向智慧图书馆的馆员能力建设[J/OL].图书馆理论与实践，2022(3):1-5.

The Intelligent Transformation of Web Archiving

Zhangxueqing

(National library of China ,Beijing 10081,China)

Abstract: The paper intends to explore the problems and challenges faced by Web archiving，and deeply analyze the current development status and difficulties of Web archiving。according to the relevant requirements of the construction of the smart library,this paper proposes the ideas and methods of the intelligent transformation of the entire project from the aspects of resource construction, data processing management, external services, etc.

Keywords: Web Archiving; Intelligent Transformation; Smart service; Knowledge service

ClC Number: G250.73

[作者简介] 张学青（1984-），男，硕士研究生，馆员，研究方向：数字资源建设，网络资源保存，智慧图书馆建设。北京，中国国家图书馆，邮编100081。电话：15120098610